Annals of Botany **98**: 647–655, 2006
doi:10.1093/aob/mcl146, available online at www.aob.oxfordjournals.org

Genetic Diversity and Structure of Managed and Semi-natural Populations of Cocoa (*Theobroma cacao*) in the Huallaga and Ucayali Valleys of Peru

DAPENG ZHANG[1,*], ENRIQUE AREVALO-GARDINI[2], SUE MISCHKE[1], LUIS ZÚÑIGA-CERNADES[2], ALEJANDRO BARRETO-CHAVEZ[2] and JORGE ADRIAZOLA DEL AGUILA[2]

[1]*USDA ARS PSI SPCL, 10300 Baltimore Avenue, Bldg 050, Rm 100, BARC-W, Beltsville, MD 20705, USA and* [2]*Intituto de Cultivos Tropiocales (ICT), Jr Santa Maria, 241, Banda del Shilcayo, Tarapoto, Peru*

Received: 31 December 2005 Returned for revision: 11 April 2006 Accepted: 30 May 2006 Published electronically: 15 July 2006

• *Background and Aims* Cocoa (*Theobroma cacao*) is indigenous to the Amazon region of South America, and it is well known that the Peruvian Amazon harbours a large number of diverse cocoa populations. A small fraction of the diversity has been collected and maintained as an *ex-situ* germplasm repository in Peru. However, incorrect labelling of accessions and lack of information on genetic diversity have hindered efficient conservation and use of this germplasm. This study targeted assessment of genetic diversity and population structure in a managed and a semi-natural population.
• *Methods* Using a capillary electrophoresis genotyping system, 105 cocoa accessions collected from the Huallaga and Ucayali valleys of Peru were fingerprinted. Based on 15 loci SSR profiles, genetic identity was examined for each accession and duplicates identified, population structure assessed and genetic diversity analysed in these two populations.
• *Key Results* Ten synonymous mislabelled groups were identified among the 105 accessions. The germplasm group in the Huallaga valley was clearly separated from the group in Ucayali valley by the Bayesian assignment test. The Huallaga group has lower genetic diversity, both in terms of allelic richness and of gene diversity, than the Ucayali group. Analysis of molecular variance suggested genetic substructure in the Ucayali group. Significant spatial correlation between genetic distance and geographical distances was detected in the Ucayali group by Mantel tests.
• *Conclusions* These results substantiate the hypothesis that the Peruvian Amazon hosts a high level of cocoa genetic diversity, and the diversity has a spatial structure. The introduction of exotic seed populations into the Peruvian Amazon is changing the cocoa germplasm spectrum in this region. The spatial structure of cocoa diversity recorded here highlights the need for additional collecting and conservation measures for natural and semi-natural cocoa populations.

Key words: *Theobroma cacao*, cocoa, conservation, germplasm, DNA fingerprinting, genetic diversity, population structure, Peru, Huallaga, Ucayali.

INTRODUCTION

Cocoa (*Theobroma cacao*) is native to the South American rainforest, but it is thought to have been domesticated in southern Mexico and the northern Central American region (Cuatrecasas, 1964; Hunter, 1990; Motamayor *et al*., 2002). The hypothesized centre of genetic diversity is located in the upper Amazonian region (Cheesman, 1944; Cuatrecasas, 1964). It is also well known that the Peruvian Amazon harbours a large number of diverse cocoa populations (Pound, 1945; Schultes, 1984; Bartley, 2005). During the past several decades, several expeditions have been made and a substantial amount of germplasm, from both wild populations and cultivated accessions, has been collected from this region (Pound, 1938, 1945; Bartley, 2005). Today, a fraction of these accessions are maintained as *ex-situ* collections in various countries (Kennedy and Mooleedhar, 1993; Lockwood and End, 1993; Motilal and Butler, 2003).

The first organized cocoa germplasm collecting expedition in the Peruvian Amazon started in 1937–1938 (Pound, 1938, 1943) and the collecting sites included Rio Nanay, Rio Morona, Rio Marañón and their tributaries. This led to the establishment of the germplasm collection in Iquitos, Peru known as the 'Pound Collection', named after the collector F. J. Pound. Many commercial clones, now called 'International clones', have their origin in this collection (Bartra, 1993; González, 1996). Pound's expeditions were aimed at searching for genotypes resistant to witches' broom disease, caused by the fungus *Crinipellis perniciosa*, and this germplasm has therefore been widely used in breeding programmes as a source of resistance to witches' broom disease.

The last major collecting expedition in the Peruvian Amazon was made in 1987–1989 in the Huallaga and

* For correspondence. E-mail ZhangD@ba.ars.usda.gov

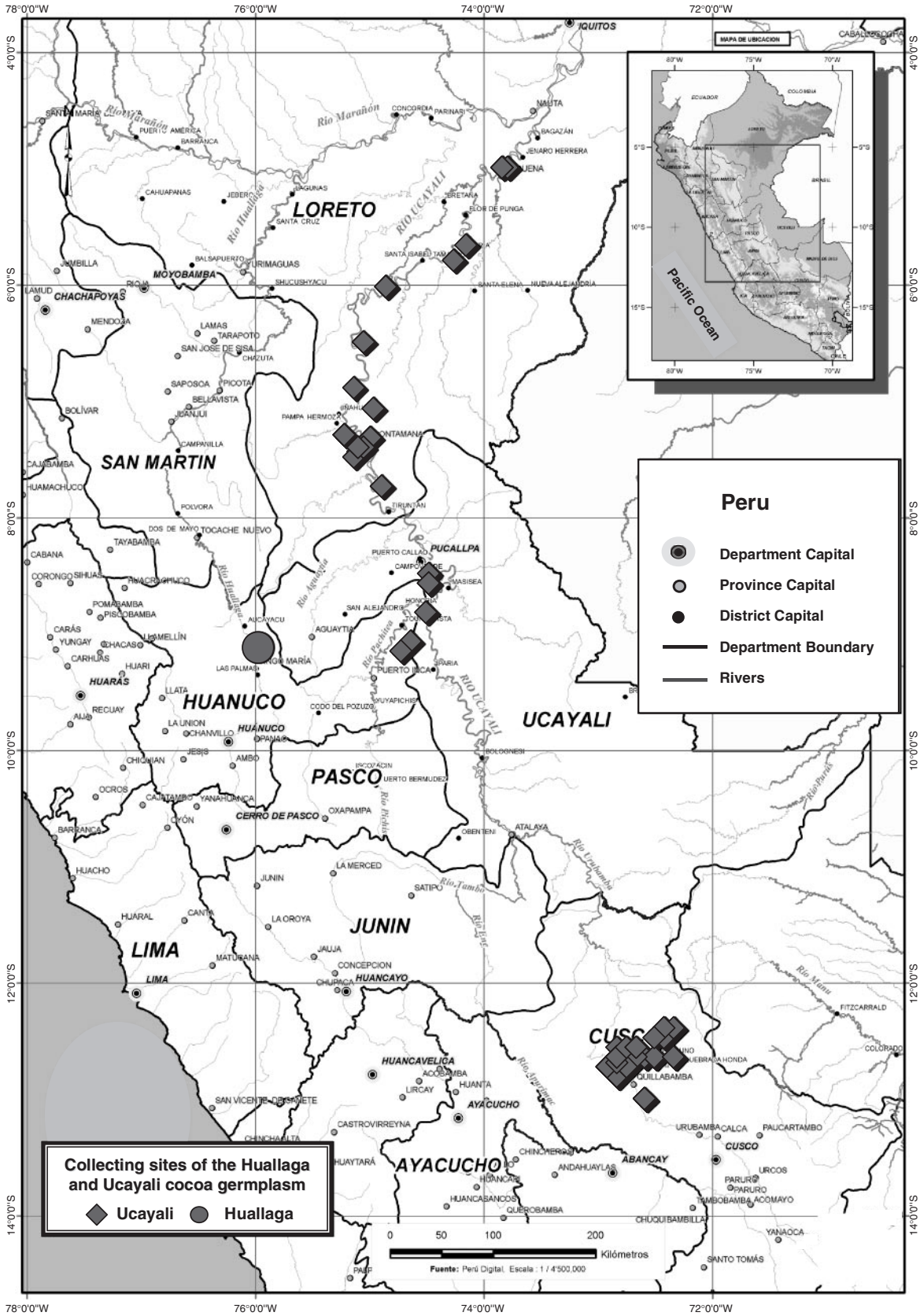

FIG. 1. Sampling sites of cocoa germplasm in the Huallaga and Ucayali valleys in Peru.

Ucayali valleys (Coral, 1988*a*; Evans *et al*., 1998). The major objective of this expedition was again to collect material with resistance to witches' broom disease. The Huallaga valley is located on the eastern slopes of the Andes. The valley runs almost south–north, to the east of the Andean Cordillera Central in San Martín and Huánuco departments of north-east Peru. The Ucayali valley is located in the east-central Peru at the foot of the Andes. It stretches from Urubamba northward to the south-west of Iquitos, including the basin both west and east of the Ucayali River, a major branch of the Amazon river (Fig. 1). These collecting activities generated two groups of well-known Peruvian germplasm, the so-called 'Huallaga clones' and the 'Ucayali clones'. The Huallaga clones were collected at 'Fundo San Jose', a cocoa farm in Naranjillo–Tingo Maria (Coral, 1988*a*, *b*). These clones were collected from cocoa fields affected by witches' broom disease (Bartra, 1993; Lopez, 1993; Rengifo, 1996). The Ucayali clones were wild cocoa trees collected from the valley of the Ucayali River as well as the Urubamba River and its tributaries (Coral, 1988*a*, *b*; Evans *et al*., 1998) including Contamana, San Carlos, Ucayali, Cushabatay and Chiatipishca Lake (Fig. 1). The Ucayali collection included two subgroups. Clones from the lower Ucayali subgroup were collected between 5°0′S and 9°10′S, whereas clones from the Urubamba subgroup were collected between 12°35′S and 13°01'S (Fig. 1). These clones were collected as budwood and were maintained in Tingo Maria. Part of the Ucayali collection was also established in Sahuayacu, near Quillabamba, Peru. However, some of the collections were lost due to the social turmoil in Peru in the following years. In 1998, re-collection was carried out in Sahuayacu to restore the Tingo Maria collection. Today, the Tingo Maria collection maintains 62 Huallaga clones and 51 Ucayali clones, as well as various other international clones (Evans *et al*., 1998). In recent years, this germplasm has increasingly attracted attention as a potential source of disease resistance. Tests for disease resistance and high yield are ongoing in Tingo Maria.

The majority of the passport information for these clones was lost during the social unrest in the late 1980s and early 1990s in Peru, and it is known that some of these clones were mislabelled. Despite their importance for local cocoa production and for cocoa breeding, this germplasm has not been systematically characterized at the molecular level. Moreover, little is known about the genetic diversity in these two groups.

Microsatellite-based DNA fingerprinting has been increasingly used in cocoa germplasm management. In recent years, this technique has been applied for individual identification (Saunders *et al*., 2004; Cryer *et al*., 2006), parentage analysis (Schnell *et al*., 2005), detection of chimaeric mutations in *in-vitro* culture (Rodriguez *et al*., 2004), diversity assessment (Lanaud *et al*., 1999, 2001) and investigation of the origin and dispersal of cocoa (Motamayor *et al*., 2002, 2003). During international forums held in England and France in 2001, a consortium of scientists and representatives from the cocoa industry, academic centres involved in cocoa research and representatives from multiple international, government-sponsored laboratories reached an agreement that a set of standardized SSR primers should be used to characterize all *T. cacao* germplasm collections (Saunders *et al*., 2001, 2004).

In this paper, the results are reported of a study in which these 15 SSR loci were used to fingerprint the 105 cocoa accessions originally collected from a managed population in the Huallaga valley and a semi-natural population in the Ucayali valley of Peru. The objective was to assess genetic identities, population structure and genetic diversity in the Huallaga and Ucayali collections. Specifically, the questions addressed included the following: (*a*) Do the Huallaga and Ucayali valleys harbour genetically distinct populations? (*b*) What is the level of genetic diversity in these two groups? (*c*) Does the cocoa genetic diversity have a spatial structure in its native habitat of the Peruvian Amazon?

This study is part of an international collaborative project on DNA fingerprinting of cocoa germplasm in Latin America. The resulting information will improve our understanding of the diversity of cocoa maintained in the Tingo Maria collection and the spatial pattern of genetic diversity in the Peruvian Amazon.

MATERIALS AND METHODS

Germplasm

Leaf samples of variable ages were collected from the cocoa germplasm collection maintained in the Universidad Nacional Agraria de la Selva, in Tingo Maria. Each tree sampled was subsequently labelled. All 62 accessions of the Huallaga clones and 43 of the 51 Ucayali clones were sampled (Table 1). The collecting localities of 105 clones are shown in Fig. 1.

DNA isolation

Theobroma cacao leaf material has high levels of endogenous phenolics that can interfere with many commercial DNA isolation procedures. Initial investigations of various DNA isolation protocols identified two methods that worked well for cocoa SSR analysis and were used interchangeably to yield consistent results. DNA was isolated from 50 mg samples of *T. cacao* leaf material using either the DNA Xtract™ Plus kit (D^2 BioTechnologies Inc., Atlanta, GA, USA) or the DNeasy® Plant System (Qiagen Inc., Valencia, CA, USA). For either method, the air-dried and frozen leaf samples were first cut into small pieces and placed in a 2-mL tube, sandwiched between ceramic spheres, with garnet matrix (Qbiogene, Carlsbad, CA, USA). Lysis solution was added following the manufacturer's recommendations, except that 10 mg mL^{-1} of polyvinylpolypyrrolidone (Sigma-Aldrich, St Louis, MO, USA) was added to the Qiagen buffer AP1. Samples were homogenized in a Bio101 Fast Prep® instrument (Qbiogene) as described previously (Saunders *et al*., 2001).

The DNA Xtract™ Plus procedure was, in brief, lysis, clarification by centrifugation, and solvent phasing followed by precipitation on ice. DNA was collected by

TABLE 1. *List of the 95 Huallaga and Ucayali clones and their assigned population membership using Bayesian clustering analysis*

Clone	Cluster	Prob.	Clone	Cluster	Probability	Clone	Cluster	Prob.
H-2	1	0·997	H-38	1	0·997	U-11	2	0·993
H-3	1	0·997	H-39	1	0·997	U-12*	2	0·450
H-4	1	0·997	H-40	1	0·997	U-15	2	0·986
H-5	1	0·998	H-41a	1	0·997	U-19	2	0·997
H-6	1	0·997	H-41	1	0·981	U-20	2	0·996
H-8	1	0·997	H-43	1	0·997	U-21	2	0·997
H-9	1	0·997	H-44	1	0·997	U-22	2	0·995
H-10	1	0·997	H-45	1	0·997	U-24	2	0·985
H-11	1	0·997	H-46	1	0·997	U-26	2	0·995
H-12	1	0·997	H-47	1	0·977	U-28	2	0·993
H-13*	1	0·727	H-48	1	0·997	U-31	2	0·879
H-15	1	0·997	H-49	1	0·997	U-32	2	0·986
H-16	1	0·996	H-50	1	0·997	U-35	2	0·996
H-18	1	0·997	H-51	1	0·997	U-36	2	0·995
H-19	1	0·997	H-52	1	0·996	U-37	2	0·997
H-20*	1	0·646	H-53	1	0·997	U-38	2	0·984
H-21	1	0·987	H-54*	1	0·461	U-39	2	0·992
H-22	1	0·997	H-55	1	0·997	U-41	2	0·957
H-23	1	0·935	H-56	1	0·997	U-43	2	0·997
H-24	1	0·997	H-57*	1	0·632	U-48	2	0·997
H-25	1	0·997	H-58*	1	0·521	U-51	2	0·994
H-26	1	0·996	H-59*	1	0·492	U-52	2	0·994
H-27	1	0·986	H-60*	1	0·516	U-53	2	0·997
H-28	1	0·997	H-61*	1	0·573	U-54	2	0·997
H-29	1	0·997	H-63	1	0·997	U-55	2	0·997
H-30	1	0·997	U-1	2	0·996	U-58	2	0·993
H-31	1	0·997	U-2*	2	0·509	U-59	2	0·993
H-32	1	0·966	U-4	2	0·988	U-65*	2	0·507
H-34	1	0·997	U-5	2	0·979	U-66	2	0·997
H-35	1	0·997	U-6	2	0·996	U-69	2	0·996
H-36	1	0·997	U-7	2	0·997	U-70	2	0·998
H-37	1	0·997	U-9*	2	0·003			

The following pairs of clones were found to be identical, and only one of the duplicate samples was included in Bayesian's clustering analysis: U-43/U-56; H-61/H-62; H-63/U-63; H-42/H-43; H-17/H-18; H-7/H-8; H-1/H-2; H-3/U-30; U-59/U-60; U-65/U-68).

* Eight Huallaga clones and five Ucayali clones had ambiguously classified membership because their assignment probability was below the criterion of 0·75. Therefore, these 13 clones were excluded in subsequent diversity analysis.

centrifugation, washed in 70 % (v/v) ethanol, centrifuged, dried and resuspended in sterile water or buffer. The DNeasy® Plant System isolation procedure included tissue lysis and RNase A treatment with 65 °C incubation, followed by centrifugation, and precipitation of detergent, proteins and polysaccharides on ice. Cell debris and precipitates were removed by centrifuging through a QIAshredder spin column assembly and the DNA in the cleared filtrate was precipitated with ethanol. This mixture was loaded onto the DNeasy column and the DNA was bound to the silica gel membrane by centrifugation. DNA was washed while bound to the membrane, and finally eluted from the membrane with preheated elution buffer. The presence of double-stranded DNA was verified by quantitation with PicoGreen® (Molecular Probes, Eugene, OR, USA) using a Fluoroskan Ascent microplate reader equipped with 485/538 excitation/emission filters (Labsystems, Helsinki, Finland).

SSR analysis

DNA amplification used primer sets with sequences previously described (Lanaud *et al*., 1999; Saunders *et al.,* 2004). Primers were synthesized by Proligo (Boulder, CO, USA) and forward primers were 5′-labelled using WellRED fluorescent dyes (Beckman Coulter, Inc., Fullerton, CA, USA). PCR was performed as described in Saunders *et al*. (2004), using commercial hot-start PCR supermixes that had been fortified with an additional 30 U of the respective hot-start *Taq* DNA polymerase (Invitrogen Platinum *Taq*, Carlsbad, CA, USA; Eppendorf HotMaster *Taq*, Brinkman, Westbury, NY, USA) added to each mL of the supermix.

The PCR products were separated by capillary electrophoresis as previously described (Saunders *et al*., 2004) using a CEQ™ 8000 genetic analysis system (Beckman Coulter Inc.). Data analysis was performed using the CEQ™ 8000 Fragment Analysis software version 7.0.55 according to manufacturer's recommendations (Beckman Coulter Inc). Fragment sizes were automatically calculated to two decimal places by the CEQ™ 8000 Genetic Analysis System. Allele calling was performed using the CEQ™ 8000 binning wizard software (CEQ™ 8000 software version 7.0.55; Beckman Coulter Inc.).

Data analysis

Duplicates in the Huallaga and Ucayali clones were assessed by identifying matching multilocus genotypes (in pairwise comparisons) among individuals. The computer program GIMLET (Valière, 2002) was used for genotype matching. Pairwise comparisons were carried out among all the 105 cocoa accessions within and among the three groups. Accessions with different names that were fully matched at 15 loci were judged to be duplicates or synonymously mislabelled accessions.

To assess the differentiation power of the 15 SSR loci, the probability of identity (PID) was calculated (Waits *et al*., 2001). The probability of identity among siblings (PID-sib), which was defined as the probability that two sibling individuals drawn at random from a population have the same mutilocus genotype, was computed (Evett and Weir, 1998; Waits *et al*., 2001).

To test if accessions from these two valleys represent genetically distinct populations, a Bayesian cluster analysis was performed using the program STRUCTURE v. 2.0 (Pritchard *et al*., 2000). A cluster assignment was defined based on the membership assignment probability. To estimate the number of subclusters (K) present in the data, STRUCTURE estimates the proportion of the genome of each individual having ancestry in each subcluster and applied a prior probability of the data [$Pr(X|K)$], where X represents the data. Here, $Pr(X|K)$ was estimated using a model allowing admixture for K from 1 to 2, because the point of this study was to assess if the two river valleys represented genetically distinct populations rather than find out how many populations were present in these samples. All STRUCTURE runs used 10 000 iterations after a burn-in of length 10 000.

To analyse the genetic diversity in the Huallaga and Ucayali collections, the intrapopulation genetic diversity was measured by estimating gene diversity (H_s) (Nei, 1987), observed heterozygosity (H_o) and F_{IS} (Wright, 1965) using GENEPOP version 3 (Raymond and Rousset, 1995). Because the two groups differed in sample size, which could affect the estimation of allelic diversity, unbiased allelic richness and private allelic richness (Leberg, 2002) were estimated using the computer program HP-Rare (Kalinowski, 2005), which performs rarefaction on measures of allelic diversity. The Exact HW test (Guo and Thompson, 1992) was used to test the deviation from Hardy–Weinberg equilibrium and was performed by GENEPOP version 3 (Raymond and Rousset, 1995).

Two methods were used to measure differences between the Huallaga and Ucayali collections. First the variation in allelic distributions between the two collections was measured. Allelic composition was assessed using a contingency table test (Weir and Cockerham, 1984) as implemented in GENEPOP version 3 (Raymond and Rousset, 1995), with a null hypothesis of identical allelic distributions in all populations. Analysis of molecular variance (AMOVA) (Excoffier *et al*., 1992) implemented in Arlequin 3.0 (Excoffier *et al*., 2005) was then used to test the significance of F_{ST} by permuting the individual genotypes between the two groups; with the probability of non-differentiation (F_{ST} not >0) being estimated over 10 000 randomizations. A modified Rogers' distance (Wright, 1978) was also calculated among all possible pairs of clones using the program TFPGA (Tools for Population Genetic Analysis) (Miller, 1997). The pairwise distances were then used to represent a Euclidean distance and were presented in a two-dimensional scaling plot using the multidimensional scaling (MDS) procedure of SAS (1999). AMOVA was also used to test if the Ucayali collection group was substructured. The difference between Urubamba and Lower Ucayali subgroups was tested using the permutation test.

To examine if there is a spatial structure in the Ucayali collection, the proportion of genetic differentiation among individuals explained by geographical distance was estimated using Mantel tests. The genetic distance between each pair of individuals in the Ucayali collection, as well as their corresponding geographical distance, was calculated. Mantel correlation tests were performed for each of the two subgroups. The Mantel test procedure in GENALEX 6 (Peakall and Smouse, 2006) was used for computation.

RESULTS

Identification of duplicates

The comparison of multilocus microsatellite profiles led to the identification of ten synonymously mislabelled groups involving 20 accessions (Table 1). The accessions within each group were defined as synonymous mislabelling because they shared exactly the same alleles across all 15 loci but were labelled with different names. With all the 15 loci considered, the combined probability of identity of sibling (PID-sib), i.e. the probability that two sibling individuals drawn at random from a population have identical genotypes, was on the order of 10^{-5} (Table 2). PID-sib is the upper limit of the possible ranges of PID in a population and thus provides

TABLE 2. *Informativeness and probability of identity (PID) of the 15 microsatellite loci, estimated from the 109 accessions in Tingo Maria collection*

Locus	Observed heterozygosity	PIC*	PID-sib/locus†	Prod(sibs)‡
Y16981	0·22	0·603	7·46E-01	7·46E-01
Y16980	0·66	0·579	3·53E-01	2·63E-01
Y16995	0·56	0·516	4·14E-01	1·09E-01
Y16996	0·54	0·405	4·31E-01	4·70E-02
Y16982	0·78	0·512	3·51E-01	1·65E-02
Y16883	0·34	0·405	4·36E-01	7·19E-03
Y16985	0·46	0·615	4·59E-01	3·30E-03
Y16986	0·54	0·757	3·55E-01	1·17E-03
Y16988	0·73	0·769	3·38E-01	3·96E-04
AJ271942	0·66	0·688	3·77E-01	1·49E-04
AJ271826	0·69	0·647	3·43E-01	5·11E-05
Y16991	0·48	0·641	4·38E-01	2·24E-05
Y16998	0·64	0·597	3·93E-01	8·81E-06
AJ271943	0·63	0·725	3·33E-01	2·94E-06
AJ271958	0·76	0·559	3·88E-01	1·14E-06
Mean	0·58	0·601		

*PIC (polymorphism information content) follows the definition of Powell *et al.* (1996).

†PID-sib (probability of identity among siblings) follows the definition of Evett and Weir (1998).

‡Accumulated PID-sib as the loci adds up, i.e. the PID-sib value of the second locus is the product of PID-sib of the first two loci.

TABLE 3. *Intrapopulation genetic diversity in Huallaga and Ucayali cocoa germplasm collection*

	Huallaga (n = 50)			Ucayali (n = 36)		
Loci	Allelic richness	Private allelic richness	Gene diversity	Allelic richness	Private allelic richness	Gene diversity
Y16981	1·695	0·224	0·102	2·356	0·884	0·464
Y16980	4·326	2·032	0·713	5·210	2·916	0·738
Y16995	2·876	0·573	0·491	5·595	3·292	0·815
Y16996	2·378	0·175	0·527	5·591	3·388	0·795
Y16982	3·955	1·039	0·712	8·613	5·697	0·905
Y16883	3·168	1·138	0·449	7·105	5·075	0·843
Y16985	2·934	0·646	0·549	6·190	3·902	0·752
Y16986	3·960	1·530	0·708	5·873	3·443	0·805
Y16988	6·554	3·681	0·848	4·536	1·663	0·587
AJ271942	3·603	1·525	0·610	5·688	3·611	0·793
AJ271826	3·856	2·741	0·689	6·293	5·179	0·780
Y16991	2·924	1·195	0·496	3·740	2·011	0·460
Y16998	3·941	1·621	0·706	5·665	3·345	0·746
AJ271943	4·581	1·779	0·736	7·868	5·066	0·881
AJ271958	4·425	1·963	0·739	5·189	2·726	0·742
Mean	3·68	1·46	0·61	5·70	3·48	0·74

Rarefaction on measures of allelic diversity (Kalinowski, 2005) was performed for the unbiased estimation of allelic richness, private allelic richness, and gene diversity.

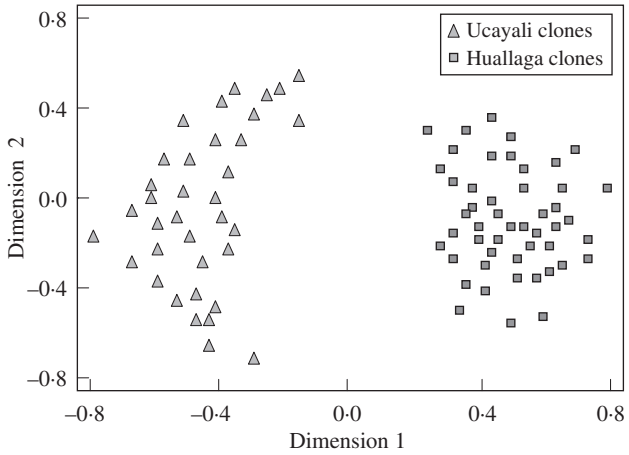

FIG. 2. Multidimensional scaling plot of 86 cocoa accessions based on Roger's distance (a representation of Euclidian distance) calculated from microsatellite data (MDS badness of fit = 0·277). All accession identifications correspond to the sample list in Table 1. Duplicated accessions and accessions with low assignment probability were excluded.

the most conservative number of loci required to resolve all genotypes. After the process of identifying duplicates, the ten accessions identified as duplicates were excluded in the subsequent analysis for population structure and genetic diversity.

Population structure and assignment test

With the prior assessment of two populations (Huallaga and Ucayali), the program STRUCTURE assigned the 95 accessions into two (or more than one) genetically inferred clusters. Each individual was associated with two probabilities, showing the degree to which its genome was classified into each cluster. The criterion for allocation was set such that when the probability of an individual of being in one cluster was >0·75, it was classified in that cluster. In other words, an individual with more than a three-quarters proportion of genetic background in the cluster should be allocated to the corresponding population, and one with less than three-quarters background in either of the two clusters should be treated as an ambiguous class member (or the 3rd population). Ambiguously classified members were not used in subsequent analyses for *F*-statistics and diversity analysis. The majority of accessions (86·9 %) could be assigned to one of the two source populations with their geographical labels correctly corresponding to their population membership. Eight Huallaga clones and five Ucayali clones had an assignment probability <75 %, and were categorized as ambiguous (Table 1).

Genetic diversities in Huallaga and Ucayali collections

All microsatellite loci were polymorphic and met the assumptions of independence (no pairs of loci were linked). A total of 161 alleles was identified, but the allelic richness differed substantially between the two groups (Table 3). The Huallaga collection had 3·7 alleles per locus, whereas the Ucayali collection had 5·7 alleles per locus. The private allelic richness was also higher in the Ucayali collection

TABLE 4. *Variation between the Huallaga and Ucayali collection and variation between the two subgroups within the Ucayali collection*

	Ucayali collection* (n = 36)			Urubamba subgroup (n = 13)
	$F_{ST}^{\dagger}$	$F_{ST}^{\ddagger}$		$F_{ST}^{\ddagger}$
Huallaga collection (n = 50)	0·207	0·234	Lower Ucayali subgroup (n = 23)	0·055
	$P < 0·001$	$P < 0·001$		$P < 0·01$

*The Ucayali collection includes 23 clones from the Lower Ucayali subgroup and 13 clones from the Urubamba subgroup. Clones in the lower Ucayali subgroup was collected from 5 °0′S to 9 °10′S, whereas clones in the Urubamba subgroup was collected from 12 °10′S to 13 °01′S.

†Definition of F_{ST} follows Weir and Cockerham (1984).

‡Definition of F_{ST} follows Excoffier *et al.* (2005). Number of permutations = 10 000.

(A = 3·5) than in the Huallaga collection (A = 1·5). Moreover, gene diversity (expected heterozygosity) in the Ucayali collection (H_e = 0·74) was substantially higher than that in the Huallaga collection (H_o = 0·61). Tests for departures from the Hardy–Weinberg equilibrium revealed numbers of diversions from the Hardy–Weinberg equilibrium in both populations, and the heterozygote deficiency was highly significant ($P < 0·001$) across populations and loci.

Significant population differentiation was detected by the contingency table test of Weir and Cockerham (1984) (F_{ST} = 0·207, $P < 0·001$). The significant divergence between the two populations was also supported by the AMOVA's permutation result (F_{ST} = 0·234, $P < 0·001$).

AMOVA showed that both the within-collection and the between-collection variations were highly significant. Twenty-one per cent of the total molecular variance was due to difference between collections, whereas 79 % was partitioned within collections. The multidimensional scaling plot showed a clear pattern of interpopulation variation (Fig. 2). With the Ucayali collection, substructure was detected using AMOVA. There was a divergence between the Urubamba subgroup and the lower Ucayali subgroup. Although the divergence is smaller than that between the Ucayali and Huallaga collections, it was statistically significant (F_{st} = 0·055, $P < 0·01$; Table 4).

Within the Ucayali collection, a moderate but significant correlation between genetic and geographical distances was detected by Mantel tests (r = 0·197, $P < 0·001$; Table 5). The influence of geography on genetic structure increased from 0·197 to 0·228 when the lower Ucayali subgroup alone was used in computation. No spatial correlation was observed in the Urubamba subgroup (r = 0·043, P = 0·46; Table 5).

DISCUSSION

Individual identification

Unambiguous identification of genotypes is a concern for cocoa germplasm management and cocoa breeding. The

TABLE 5. *Mantel test for correlation between genetic and geographical distances in the Ucayali collection and the two subgroups*

		SS*X**	SS*Y*†	SP*XY*‡	R*XY*§	Probability
Ucayali collection	$n = 36$	48308·2	80474368·7	387458·9	0·197	0·001
I. Urubamba subgroup	$n = 13$	19515·4	6564184·4	81593·9	0·228	0·018
II. Lower Ucayali subgroup	$n = 23$	6457·4	54431·7	340·9	0·018	0·397

* Sum of products of the *X* matrix (genetic distance) elements.
† Sum of products of the *Y* matrix (geographical distance) elements.
‡ Sum of cross products of corresponding elements of the *X* and *Y* matrices.
§ Mantel correlation coefficient.

cocoa trees in the various collections were obtained at different times with limited information about their correct identity. Genotypes can be difficult to distinguish morphologically and identification relies heavily on plant labels and field maps. Over the years, a significant proportion of accessions were mislabelled, or their genetic identity is not clear (Motilal and Butler, 2003; Turnbull *et al*., 2004). An agreement among the various laboratories was reached to use 15 standardized SSR primers to characterize all *T. cacao* germplasm collections (Saunders *et al*., 2004). In the present study, it has been demonstrated that this set of SSR primers was effective for the assessment of genetic identity of cocoa germplasm.

However, because some of the Ucayali clones in the Tingo Maria collection were lost during the social unrest in the late 1980s, and only a fraction of these accessions were re-collected from Sahuayacu and other locations in recent years, the collection of 'original living trees' needed as reference standards is not complete. Therefore, assessment of genetic identity in this study was limited with regard to duplicate identification and assignment of individuals to their source populations. The SSR fingerprint profiles demonstrated that each accession is a unique genotype, which can be correctly assigned to its source population.

Genetic diversity and population structure

The overall genetic diversity was high in the two germplasm collections, compared with the previously published studies in cocoa. The average number of alleles per locus based on rarefaction measurement was 5·7 in the Ucayali collection, which is comparable to the allele richness in a diverse set of cocoa (140 accessions with different geographical origins) maintained in the USDA Mayaguez Research Station in Puerto Rico (Zhang *et al*., 2006). The allelic richness found in the present study is comparable with our unpublished diversity survey (D. Zhang, M. Boccara and D. Butler) in neighbouring river valleys in Peru, including the germplasm in the valleys of Rio Nanay, Rio Morona and Rio Marañón. This finding is also in agreement with the wide range of morphological variation observed in this germplasm (Coral, 1988*a*; Evans *et al*., 1998). Peruvian Amazon is believed to be a centre of diversity for the genus *Theobroma* (Bartley, 2005). Cuatrecasas (1964) listed seven species from this region. The results provide more evidence substantiating the hypothesis that the Peruvian Amazon hosts a high level of genetic diversity of *T. cacao*.

The higher allele richness in the Ucayali collection than in the Huallaga collection could be explained by the fact that the Ucayali clones were collected from a much wider geographical region. The Ucayali clones were collected as different subgroups over a 2-year period (1987–1989) at different geographical locations, ranging from Pucalpa to Quillabamba. Most of the Huallaga clones, on the other hand, were collected in one cocoa growing area in Naranjillo (Fig. 1). Moreover, as a cultivated species, the diversity level of cocoa populations is subject to strong human intervention. The Huallaga clones sampled in Naranjillo were no longer representative of natural or semi-natural populations. The diversity level and distribution described is probably a result of natural forces and human intervention combined.

Gene flow is a critical parameter for understanding the process of species dispersal and local adaptations. Microsatellite markers, in combination with spatial statistical tools, offer an indirect method to measure gene flow. In species with restricted gene flow, a pattern of 'isolation by distance' is expected because the genetic distances among individuals are positively correlated with geographical distance (Wright, 1943; Rousset, 1997). As a Neotropical tree species, cocoa was assumed to have restricted gene flow, due to its limited distance of seed dispersal by rodents, insect-mediated dispersal of pollen, and the large spatial distances separating patches in the Amazon rainforest. So far, little information is available regarding spatial pattern and gene flow in cocoa. In the present study, a moderate but significant spatial correlation was detected in the Ucayali population (Table 5). However, in the Urubamba subgroup, no spatial correlation was observed (Table 5). This disagreement could be explained by the collecting localities of the two subpopulations. In the case of the Ucayali subpopulation, the collecting localities stretched over a large north–south latitude gradient, ranging from 5 °0′S to 9 °10′S.

This correlation with geographic distance was confounded with other selection forces, such as climatic and soil parameters, which also have a trend of latitude changes. In the case of the Urubamba subpopulation, all the accessions were collected within a limited latitude range, where the small latitude gradient may not play a role of selection for local adaptation.

Thus, the spatial pattern of genetic diversity detected in the Ucayali population only provided circumstantial evidence to support the notion of an isolation-by-distance influence on gene flow. Further studies including different

spatial scales of sampling would allow a verification of the north–south spatial autocorrelation with possible climatic and soil parameters.

Implications for conservation and breeding

The present results substantiate the hypothesis that the Peruvian Amazon hosts a high level of genetic diversity, and the diversity has a spatial structure in the native habitat of cocoa. The introduction of breeding progenies in the rehabilitation programme is changing the cocoa germplasm spectrum in this region. Identifying the patterns of distribution of intraspecific genetic variation can provide data concerning the temporal and special dynamics of this economically important crop, which has not been previously studied in depth at the population level. The spatial structure of cocoa diversity recorded here highlights the need for additional collecting and conservation measures for natural and semi-natural cocoa populations in the Peruvian Amazon.

The three main cocoa diseases, witches' broom, frosty pod rot and black pod, constitute a serious threat to the livelihoods of cocoa farmers in Peru and Latin America in general. Cocoa production in the Americas has dropped by 75 % in last 16 years largely due to these three diseases. During the past several decades, large numbers of hybrid seeds were introduced into this region, especially in the Huallaga valley. These hybrid seeds were progenies derived from crosses among the International clones from Pound's collection, as well as other selected cocoa clones introduced from Central America and Caribbean countries, i.e. the UF (United Fruit) and ICS (Imperial College Selections) clones, to produce hybrid seeds for this region (Evans *et al*., 1998). In recent years, cocoa rehabilitation has been expanding in Peru. Approximately 16 000 ha of abandoned plantations have the potential for rehabilitation, and there are over 200 000 ha of land suitable for cocoa production in the Amazonas Department alone (Fuell, 2003). Thus Peru has the potential to be a significant producer in the long term, if disease problems can be addressed and the security situation stabilizes sufficiently. As elite cocoa germplasm is re-introduced into this region, and there is increasing forest fragmentation, the change of temporal and spatial distribution of cocoa diversity in Huallaga and Ucayali will continue. The results thus provide useful baseline data for monitoring future changes in these valleys.

ACKNOWLEDGEMENTS

We thank Stephen Pinney and Eric Tillson for their contributions to the genotyping, and Tulio Pisco-Rojas, Henry Yalta, Katerine Rengifo, Janet Gonzales, Cecilia Ortiz and Kennet Reategui for their technical support in collecting samples from the field genebank. We also thank Universidad Nacional Agraria de la Selva of Tingo Maria for help with sample collecting. Special thanks are due to Ainong Shi, Lambert Motilal, Ranjana Bhattacharjee and Vishnarayan Mooleedhar who reviewed the manuscript and made critical suggestions for revision. This work was supported in part by the INCAGRO/MINAG project. Funding to pay the Open Access publication charges for this article was provided by the USDA.

LITERATURE CITED

Bartley BGD. 2005. *The genetic diversity of cacao and its utilization.* Wallingford, UK: CABI Publishing.

Bartra TE. 1993. *Caracterización botánica-agronómica ex situ de 20 clones de cacao (*Theobroma cacao *L.) colectados en la cuenca del río Huallaga.* Thesis Ing Agr, Universidad Nacional Agraria de la Selva, Tingo Maria, Peru.

Cheesman EE. 1944. Notes on the nomenclature, classification and possible relationships of cocoa populations. *Tropical Agriculture* **21**: 144–159.

Coral F. 1988*a*. *Expedicão ão Vale do Rio Ucayali (1987–1988)*. Field notes and report of Cacao Germplasm Collections from the Rio Ucayali drainage system in Peru.

Coral F. 1988*b*. *Desarrollo de la producción y procesamiento de cacao en la Region de Tingo Maria.* Informe Tecnico. Proyecto FD/PER/86/458. OPS/PNUD/UNFDAC. Available from Instituto de Cultivos Tropicales, San Martin, Peru (ict@terra.com.pe)

Cryer NC, Fenn MGE, Turnbull CJ, Wilkinson MJ. 2006. Allelic size standards and reference genotypes to unify international cocoa (*Theobroma cacao* L.) microsatellite data. *Genetic Resource and Crop Evolution* (http://dx.doi.org/10.1007/s10722-005-1286-9).

Cuatrecasas J. 1964. Cacao and its allies. A taxonomic revision of the genus *Theobroma*. *Contributions from the United States National Herbarium* **35**: 375–614. Washington, DC: Smithsonian Institution Press.

Evans HC, Krauss U, Rutz RR, Acosta TZ, Arevalo-Gardini E. 1998. Cocoa in Peru. *Cocoa Growers Bulletin* **51**: 7–22.

Evett IW, Weir BS. 1998. *Interpreting DNA evidence: statistical genetics for forensic scientists*. Sunderland, MA: Sinauer.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**: 47–50.

Fuell LD. 2003. *Peru cocoa production and outlook 2003.* Foreign Agricultural Service GAIN Report #PE3006, USDA

González MT. 1996. *Caracterización botánico-agronómico de 25 clones internacionales de cocoa (*Theobroma cocoa *L).* Thesis Ing Agr, Universidad Nacional Agraria de la Selva, Tingo Maria, Peru.

Guo SW, Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics* **48**: 361–72.

Hunter RJ. 1990. The status of cocoa (*Theobroma cacao*, Sterculiaceae) in the western hemisphere. *Economic Botany* **44**: 425–439.

Kalinowski ST. 2005. HP-Rare: a computer program for performing rarefaction on measures of allelic diversity. *Molecular Ecology Notes* **5**: 187–189.

Kennedy AJ, Mooleedhar V. 1993. Conservation of cocoa in field genebanks—the International Cocoa Genebank, Trinidad. In: *Proceedings of the International Workshop on Conservation, Characterization and Utilization of Cocoa Genetic Resources in the 21st Century*. Port-of-Spain, Trinidad and Tobago: The University of the West Indies, Cocoa Research Unit, 21–23.

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL. 1999. Isolation and characterization of microsatellites in *Theobroma cacao* L. *Molecular Ecology* **8**: 2141–2143.

Lanaud C, Motamayor JC, Risterucci AM. 2001. Implications of new insight into the genetic structure of *Theobroma cacao* L. for breeding strategies. In: *Proceedings of the International Workshop on New Technologies for Cocoa Breeding*, Kota Kinabalu, Malaysia. . . . London: Ingenic Press, 89–107. http://www.personal.psu.edu/users/a/o/aoa113/ingenic/documents/communications/meetings/past/2000INGENIC.pdf (30 December 2005).

Leberg PL. 2002. Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology* **11**: 2445–2449.

Lockwood C, End M. 1993. History, technique and future needs for cacao collection. In: *Proceedings of the International Workshop on Conservation, Characterization and Utilization of Cocoa Genetic Resources in the 21st Century*. Port-of-Spain, Trinidad and Tobago: The University of the West Indies, Cocoa Research Unit, 1–14.

Lopez H. 1993. *Caracterización botánica-agronómica de 20 clones de cacao (Theobroma cacao* L.*) recolectados en las cuencas de los ríos Ucayali y Urubamba.* Thesis Ing Agr, Universidad Nacional Agraria de la Selva, Tingo Maria, Peru.

Miller M. 1997. TFPGA—Tools for population genetic analyses, version 1·3, Northern Arizona University. http://iubio.bio.indiana.edu:7780/archive/00000446/ (30 December 2005)

Motamayor JC, Lopez PA, Ortiz CF, Moreno A, Lanaud C. 2002. Cacao domestication. I. The origin of the cacao cultivated by the Mayas. *Heredity* **89**: 380–386.

Motamayor JC, Risterucci AM, Heath M, Lanaud C. 2003. Cacao domestication. II. Progenitor germplasm of the Trinitario cacao cultivar. *Heredity* **91**: 322–330.

Motilal L, Butler D. 2003. Verification of identities in global cacao germplasm collections. *Genetic Resources and Crop Evolution* **50**: 799–807.

Nei M. 1987. *Molecular evolutionary genetics*. New York, NY: Columbia University Press.

Peakall R, Smouse PE. 2006. Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**: 288–295

Pound FJ. 1938. Cacao and witchbroom disease (*Marasmius perniciosus*) of South America. *Archives of Cocoa Research* **1**: 20–72.

Pound FJ. 1943. *Cacao and witches' broom disease (*Marasmius perniciosus*).* Report on a recent visit to the Amazon territory of Peru, September, 1942–February, 1943. Yuille's Printery, Port of Spain, Trinidad and Tobago.

Pound FJ. 1945. A note on the cocoa population of South America. In: *Report and Proceedings of the 1945 Cocoa Conference*, London, 131–133.

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, *et al*. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding* **2**: 225–238.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure from multilocus genotype data. *Genetics* **155**: 945–959.

Raymond M, Rousset F. 1995. GENEPOP (version 1.3) population genetic software for exact tests and ecumenicism. *Journal of Heredity* **86**: 248–249. http://wbiomed.curtin.edu.au/genepop/ (30 December 2005)

Rengifo KE. 1996. *Caracterización botánica-agronómica de 14 clones de cacao de la colección Huallaga del Banco de Germoplasma de cacao (*Theobroma cacao *L.) en Tingo Maria.* Thesis Ing Agr, Universidad Nacional Agraria de la Selva, Tingo Maria, Peru.

Rodriguez LCM, Wetten AC, Wilkinson MJ. 2004. Detection and quantification of *in vitro*-culture induced chimerism using simple sequence repeat (SSR) analysis in *Theobroma cacao* (L.). *Theoretical and Applied Genetics* **110**: 157–166.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**: 1219–1228

SAS. 1999. *SAS Version 8.02: SAS/STAT Software: changes and enhancements through Release 8.02*. Cary, NC: SAS Institute Inc.

Saunders JA, Hemeida AA, Mischke S. 2001. USDA DNA fingerprinting programme for identification of *Theobroma cacao* accessions. In: *Proceedings of the International Workshop on New Technologies for Cocoa Breeding*, Kota Kinabalu, Malaysia. London: Ingenic Press, 108–114. http://www.personal.psu.edu/users/a/o/aoa113/ingenic/documents/communications/meetings/past/2000INGENIC.pdf (30 December 2005)

Saunders JA, Mischke S, Leamy EA, Hemeida AA. 2004. Selection of international molecular standards for DNA fingerprinting of *Theobroma cacao*. *Theoretical and Applied Genetics* **110**: 41–47.

Schnell RJ, Olano CT, Brown JS, Meerow AW, Cervantes-Martinez C, Nagai C, *et al*. 2005. Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. *Journal of the American Society of Horticultural Science* **130**: 181–190.

Schultes RE. 1984. Amazonian cultigens and their northward and westward migrations in pre-Columbian times. In: Stone D, ed. *Pre-Columbian plant migration*. Cambridge, MA: Harvard University Press, 32–33.

Turnbull CJ, Butler DR, Cryer NC, Zhang D, Lanaud C, Daymond AJ, *et al*. 2004. Tackling mislabelling in cocoa germplasm collections. *INGENIC Newsletter* **9**: 8–11.

Valière N. 2002. Gimlet, a computer program for analysing genetic individual identification data. *Molecular Ecology Notes* **2**: 377–379.

Waits LP, Luikart G, Taberlet P. 2001. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology* **10**: 249–256.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Wright S. 1943. Isolation by distance. *Genetics* **28**: 114–138.

Wright S. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19**: 395–420.

Wright S. 1978. *Evolution and the genetics of populations.* Vol. 4. *Variability within and among natural populations.* Chicago, IL: University of Chicago Press.

Zhang D, Mischke S, Goenaga R, Hemeida AA, Saunders JA. 2006. Accuracy and reliability of high-throughput microsatellite genotyping for cacao clone identification. *Crop Science* (in press)